

03 020
YOR.

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2003年 2月27日

出 願 番 号

Application Number:

特願2003-050916

[ST.10/C]:

[JP 2003-050916]

出 願 人

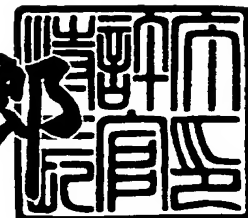
Applicant(s):

インターナショナル・ビジネス・マシーンズ・コーポレーション

2003年 5月20日

特 許 庁 長 官
Commissioner,
Japan Patent Office

太田 信一郎



出証番号 出証特2003-3037030

【書類名】 特許願

【整理番号】 JP9030020

【提出日】 平成15年 2月27日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/00
G06F 17/22

【発明者】

【住所又は居所】 神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ピー・エム株式会社 東京基礎研究所内

【氏名】 渋谷 哲朗

【特許出願人】

【識別番号】 390009531

【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】 100086243

【弁理士】

【氏名又は名称】 坂口 博

【代理人】

【識別番号】 100091568

【弁理士】

【氏名又は名称】 市位 嘉宏

【代理人】

【識別番号】 100108501

【弁理士】

【氏名又は名称】 上野 剛史

【復代理人】

【識別番号】 100110607

【弁理士】

【氏名又は名称】 間山 進也

【手数料の表示】

【予納台帳番号】 062651

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9706050

【包括委任状番号】 9704733

【包括委任状番号】 0207860

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 塩基配列のスクリーニングを行うためのコンピュータ・システム、そのための方法、該方法をコンピュータに対して実行させるためのプログラムおよび該プログラムを記憶したコンピュータ可読な記録媒体

【特許請求の範囲】

【請求項1】 塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システム。

【請求項2】 前記コンピュータ・システムは、前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納する最大編集距離格納部を含む、

請求項1に記載のコンピュータ・システム。

【請求項3】 前記評価処理部は、評価停止を動的に判断する停止判断部を含み、前記停止判断部は、前記最大編集距離を越えて前記相補配列データの評価が行われることを判断する、

請求項2に記載のコンピュータ・システム。

【請求項4】 塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部

分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システム。

【請求項 5】 前記コンピュータ・システムは、前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納する最大編集距離格納部を含む、

請求項 4 に記載のコンピュータ・システム。

【請求項 6】 前記評価処理部は、評価停止を動的に判断する停止判断部を含み、前記停止判断部は、前記最大編集距離を越えて前記相補配列データの評価が行われることを判断する、

請求項 5 に記載のコンピュータ・システム。

【請求項 7】 前記評価処理部は、前記動的停止判断部の判断に応答して、前記部分配列指定部に対して直前に評価を実行していた部分配列と所定の重なり合いを有する他の部分配列を指定させる、

請求項 5 に記載のコンピュータ・システム。

【請求項 8】 コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法。

【請求項 9】 前記評価ステップは、さらに前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断結果に応答して停止信号を生成するステップと

を含む請求項 8 に記載のコンピュータ制御方法。

【請求項 10】 コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法。

【請求項 11】 前記評価ステップは、少なくとも前記部分配列のうちの 1 つについて前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断の結果に応答して動的に評価を停止するステップとを含む、

請求項 10 に記載のコンピュータ制御方法。

【請求項 12】 前記評価ステップは、さらに評価の停止を判断するステップと、

前記評価の停止の判断に応答して、直前に評価を実行していた部分配列と所定の重なり合いを有する他の部分配列を指定するステップと

を含む、請求項 11 に記載のコンピュータ制御方法。

【請求項 13】 コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、前記コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラム。

【請求項 14】 前記評価ステップは、さらに前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断結果に応答して評価を停止させるステップと

を含む請求項 13 に記載のコンピュータ実行可能なプログラム。

【請求項 15】 コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラム。

【請求項 16】 前記評価ステップは、少なくとも前記部分配列のうちの 1 つについて前記最大編集距離を越えて前記相補配列データの評価が行われることを判断するステップと、

前記判断の結果に応答して動的に評価を停止するステップと、

前記評価の停止の判断に応答して、直前に評価を実行していた部分配列と所定の重なり合いを有する他の部分配列を指定するステップとを含む、請求項 15 に記載のコンピュータ実行可能なプログラム。

【請求項 17】 請求項 13 に記載のプログラムを記録したコンピュータ可読な記憶媒体。

【請求項 18】 請求項 15 に記載のプログラムを記録したコンピュータ可読な記憶媒体。

【請求項 19】 塩基配列のスクリーニングをネットワークを介して実行するためのサーバであって、該サーバは、

ターゲット塩基配列データを格納するデータベースと、

ネットワークを介してターゲット指定データとターゲット塩基配列に結合する可能性のあるプローブ塩基配列を受け取ると共に、スクリーニングの結果を前記ネットワークを介して送出する送受信部と、

前記プローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット指定データを使用して前記データベースを検索し、検索されたターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行すると共に、前記送受信部に伝送する評価処理部と、

を含むサーバ。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、塩基配列のスクリーニングに関し、より詳細には所定の塩基配列が、他の塩基配列の一部分を構成するか否かを効率よく判定するために、所定の塩基配列をスクリーニングするためのシステム、そのための方法、該方法をコンピュータに対して実行させるためのプログラム、該プログラムを記憶したコンピュータ可読な記録媒体およびネットワークを介して塩基配列のスクリーニングを実行させるためのサーバに関する。

【 0 0 0 2 】

【従来の技術】

DNAは、A（アデニン）、T（チミン）、C（シトシン）、G（グアニン）からなる4種類の塩基から構成される塩基鎖であり、RNAはA、U（ウラシル）、G、Cからなる4種類の塩基から構成されることが知られている。これらの合計5種類の塩基は、相補塩基と呼ばれる組み合わせでそれぞれ特異的に結合することが知られており、アデニン（A）とチミン（T）（RNAの場合にはアデニン（A）とウラシル（U））、シトシン（C）とグアニン（G）とが、それぞれ結合し、2重鎖を形成することが知られている。この2重鎖がハイブリダイゼーションとして参照されており、2重鎖を生成することをハイブリダイズする、として参照されている。また、所与の塩基配列に対して、2重鎖を形成するように結合することが可能な塩基配列は、相補配列として参照されている。

【 0 0 0 3 】

上述した各塩基間の相補性という性質を利用することにより、所定のDNAや、RNAを含む可能性のある試料中において、目的とするDNAやRNA（以下、本発明においては、被検査対象とされる塩基配列のことを、ターゲット塩基配列として参照する。）が存在するか否か、を判断することが必要とされる場合がある。さらに、本発明においてはプローブとして参照され、DNAやRNAの所与の領域の塩基に対して特異的に結合する塩基鎖を用意し、この塩基鎖がターゲット配列に対して結合、すなわちハイブリダイゼーションするかどうかを判断する必要が生じることもある。上述した判断を、結合性判断またはハイブリダイゼーション判断として参照する。また、ターゲット塩基配列を検出するために使用する相補配列をプローブとして参照する。

【0004】

上述した結合性の判断は、種々の目的のために利用されており、例えばDNAチップは、上述したハイブリダイゼーションの性質を利用する情報処理チップである。情報処理チップは、多くの場合様々なDNA、RNAといった塩基配列に対する相補配列を用意し、同時に大量のハイブリダイゼーション判断を行わせることにより、目的とする処理を実行する。また、いわゆるPCR (Polymerase Chain Reaction) は、DNAの配列上の2箇所の部分の配列に対して相補配列を作成し、これらの相補配列で囲まれる領域を大量にコピーすることにより、DNAの配列に対する判断・評価を行う手法である。

【0005】

これらの相補配列を使用する判断・評価において多くの場合、対象とされる現実の試料に目的とは異なる別のDNAやRNAの配列が混在する場合がある。このような場合、作成または提供されるプローブが、混在する塩基配列には結合しないことを保証することにより、判断・評価の効率化・高精度化、および高信頼化を達成することが可能となる。また、所定のDNAに対して特異的なプローブPを作成するためにDNAやRNAのシンセサイザが使用される場合もあり、目的以外のプローブPを効率よく排除してゆくことにより、タンパク合成およびスクリーニングを含めた効率は、著しく高められるものと考えられる。

【0006】

図18は、ターゲット塩基配列と、プローブとの関係を示した図である。図18においては、ターゲット塩基配列が符号Tで示されており、プローブが符号Pで示されている。ターゲット塩基配列Tは、例えば数千の塩基が結合した長鎖の塩基配列とすることができる（以下、本発明においては塩基数のことを、base pair: bpとして記述する。）。本質的には図18に示されるプローブ102は、ターゲット塩基配列Tのうち、T_Pで示される領域の配列の完全な相補配列となっていることが理想的である。

【0007】

しかしながら、所定の塩基配列と別の塩基配列とは、現実的には完全な相補配列でなくても結合する性質がある（80%－90%の一致でも結合する）。また

、シーケンサなどの塩基配列決定装置における分析誤差もあるので、必ずしも100%完全に一致しないからといって、プローブの候補となり得ないとして排除することは妥当ではない。このため、プローブPとターゲット塩基配列とが結合しないことを確認するためには、従来は、Smith-Waterman法などの高精度アライメント・アルゴリズムを適用して相互の塩基配列の解析を行い、プローブの相補配列に近い配列がターゲット塩基配列に全くないことを保証する必要とされていた。

【0008】

また、これまで相互に近似する塩基配列を検索する手法として、BLAST (Altschul SF., Miller, GW., Myers EW., Lipman DJ., "Basic local alignment search tool", J. Mol. Biol. 1990, Oct. 5, 215 (3), 403-410)、FASTA (Pearson, WR., Lipman, DJ., "Improved tools for biological sequence comparison", Proc. Natl. Acad. Sci. USA, 1988, Apr; 85(8), 2444-2448 Related Articles, Links)などの高速検索アルゴリズムが提案されている。しかしながら、これらのアルゴリズムでも上述したようなターゲット塩基配列中における類似する部分配列を全部発見できるわけではなく、ターゲット配列に対してプローブが結合しないという保証を行うスクリーニング用途には適するものではない。

【0009】

【発明が解決しようとする課題】

本発明は、従来技術の上述した不都合に鑑み、プローブPが与えられた場合に、プローブPとターゲット塩基配列Tとが少なくとも結合しないという評価の結果を、ユーザに対して迅速に提供することを可能とするものである。すなわち、プローブPとターゲット塩基配列Tとが少なくとも結合しないことをユーザに対して示すことができれば、迅速にプローブの有効性および信頼性についての情報をユーザに対して提供することが可能となる。これまで種々の検索手法が提案されていたものの、ユーザの許容する範囲の誤差で所与のターゲット塩基配列に対して結合しないこと、または結合するという情報に加え、結合する場合には、可能性のあるプローブPの配列をユーザに対して提示することを可能とする塩基配列のスクリーニング・システム、そのための方法、該方法をコンピュータに対し

て実行させるためのプログラム、該プログラムを記録したコンピュータ可読な記憶媒体、およびそのためのサーバが必要とされていた。

【0010】

【課題を解決するための手段】

本発明は、ユーザが要求する精度内ではプローブPとターゲット塩基配列Tとが少なくとも結合しないことをおよびユーザが指定する許容範囲内で互いに結合する可能性がある場合には、結合する可能性と、相似度および該当するサブシーケンスの情報をユーザに対して提示することができれば、高速なプローブのスクリーニングが可能になるという着想のもとになされたものである。

【0011】

本発明の上記課題を解決するために、本発明においては、ユーザが最大の編集距離を指定し、最大編集距離だけ相違するか否かの評価から評価を始め、降順にプローブPとターゲット塩基配列Tとの結合性についての評価を行う。その結果、ユーザが指定する最大編集距離よりも編集距離が大きくなることが明らかになった時点でターゲット塩基配列全体の評価を行わずに「結合しない」ことについての情報を提供することができる。さらに、本発明においては、評価停止条件をサブシーケンスと相補配列Qとの間の塩基差に対応させて動的に変更させることにより、結合する可能性があると判断された場合には、それらの相似度、塩基配列といった付帯的データをユーザに対して提供することが可能となる。

【0012】

さらに、本発明においてはさらに精度を向上させ、より長鎖のターゲット塩基配列に対しても高速化、高精度化を確実に提供することを可能とする。すなわち、本発明の他の態様では、ターゲット塩基配列は、一定の重複を持つ部分配列 T_p に分割され、部分配列 T_p ごとにプローブPと結合しないことを判定する。より具体的には、評価したいターゲット塩基配列Tを適当なオーバーラップを与えた複数の部分配列 T_p ($p = 1 \sim w$) に分割する。分割された部分配列 T_p ごとに、プローブPの配列の相補配列が部分配列 T_p に含まれるか否かの判断を実行することで、さらに信頼性を高めることが可能となる。また、ターゲット塩基配列を部分配列 T_p に分割する場合の鎖長について、本発明の方法を使用すること

ができる。

【 0 0 1 3 】

すなわち、本発明によれば、塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システムが提供される。

【 0 0 1 4 】

本発明の前記コンピュータ・システムは、前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納する最大編集距離格納部を含むことが好ましい。

【 0 0 1 5 】

本発明の前記評価処理部は、評価停止を動的に判断する停止判断部を含み、前記停止判断部は、前記最大編集距離を越えて前記相補配列データの評価が行われることを判断することができる。

【 0 0 1 6 】

本発明によれば、塩基配列のスクリーニングを実行するためのコンピュータ・システムであって、該システムは、

ターゲット塩基配列データを格納するターゲット塩基配列格納部と、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プロ

ープ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行する評価処理部と、

前記評価処理部により得られた評価結果を格納する記憶部と

を含む塩基配列のスクリーニングを実行するためのコンピュータ・システムが提供される。

【 0 0 1 7 】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ制御方法が提供される。

【 0 0 1 8 】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法であって、該方法は、コンピュータ・システムに対して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部

分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと
 を実行させる、コンピュータ制御方法が提供される。

【0019】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、前記コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと
 を実行させる、コンピュータ実行可能なプログラムが提供される。

【0020】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるためのコンピュータ制御方法のためのコンピュータ実行可能なプログラムであって、該プログラムは、コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部

分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ実行可能なプログラムが提供される。

【 0 0 2 1 】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるコンピュータ制御方法のためのコンピュータ実行可能なプログラムが記録されたコンピュータ可読な記憶媒体であって、該プログラムは、前記コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集距離を格納するステップと、

前記各塩基配列データおよび前記最大編集距離を各格納部から読み出して前記ターゲット塩基配列データと前記相補配列データとを編集距離の大きな順から結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ可読な記憶媒体が提供される。

【 0 0 2 2 】

本発明によれば、コンピュータ・システムに対して塩基配列のスクリーニングを実行させるコンピュータ制御方法のためのコンピュータ実行可能なプログラムが記録されたコンピュータ可読な記憶媒体であって、該プログラムは、コンピュータ・システムを制御して、

ターゲット塩基配列データおよびプローブ塩基配列を格納するステップと、

ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成して格納するステップと、

前記ターゲット塩基配列と前記プローブ塩基配列との間に許容する最大の編集

距離を格納するステップと、

前記ターゲット塩基配列データから短鎖の部分配列を指定すると共に、前記部分配列ごとに前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を評価するステップと、

前記評価の結果を記憶部に格納するステップと

を実行させる、コンピュータ可読な記憶媒体が提供される。

【 0 0 2 3 】

本発明によれば、塩基配列のスクリーニングをネットワークを介して実行するためのサーバであって、該サーバは、

ターゲット塩基配列データを格納するデータベースと、

ネットワークを介してターゲット指定データとターゲット塩基配列に結合する可能性のあるプローブ塩基配列を受け取ると共に、スクリーニングの結果を前記ネットワークを介して送出する送受信部と、

前記プローブ塩基配列から相補配列データを生成して格納する相補配列データ格納部と、

前記ターゲット指定データを使用して前記データベースを検索し、検索されたターゲット塩基配列データと前記相補配列データとを編集距離の大きな順に評価して、前記プローブ塩基配列の前記ターゲット塩基配列に対する結合可能性を判断する処理を実行すると共に、前記送受信部に伝送する評価処理部と、

を含むサーバが提供される。

【 0 0 2 4 】

【発明の実施の形態】

以下、本発明につき図面に示した具体的な実施の形態をもって説明するが、本発明は、後述する実施の形態に限定されるものではない。本発明における結合性判断において、ターゲット塩基配列から得られるコンピュータ処理可能なターゲット塩基配列データを、以下、ターゲット塩基配列Tとして参照し、プローブである塩基配列から得られるコンピュータ処理可能な塩基配列データを、プローブPとして参照する。さらに、プローブPのコンピュータ処理可能な相補配列デー

タを相補配列Qとして参照する。

【0025】

A：本発明における使用パラメータおよび本質的処理の概説

図1は、ターゲット塩基配列Tと、プローブPと、相補配列Qとの関連性を示した図である。図1(a)に示すように、本発明の特定の実施の形態において、プローブPを $P = \text{ATGGCCTGAC}$ とすると、Pの相補配列Qは、 $Q = \text{GT CAGGCCAT}$ となる。図1(a)に示すようにプローブPと相補配列Qとの関係は、各塩基を相補塩基に置き換えただけではなく、配列の向きが図1(a)に示される矢線ALにしたがい、逆となっていることに留意されたい。また、相補配列Qは、ユーザが指定する最大編集距離内においてターゲット塩基配列Tの一部を構成する配列部分となっていることが、本発明の評価の本質的な処理となる。

【0026】

このとき、本発明においてはターゲット塩基配列Tと、プローブPまたは相補配列Qとの間に相似度 r を定義する。本発明においては、ターゲット塩基配列TとプローブPとが結合するために、少なくともターゲット塩基配列Tが、プローブPとの相似度で r 以上の配列部分を含むことを要求する。すなわち、本発明では、ターゲット塩基配列TとプローブPとは、ターゲット塩基配列中に相似度 r の配列部分がなければ、結合できないものとされる。

【0027】

相似度 r を決定するためには種々の方法を使用することができ、本発明において具体的には、以下の基準を使用して相似度 r を決定することができる。ターゲット塩基配列Tに含まれるサブシーケンスSとすると、サブシーケンスSから相補配列Qを生成するために、塩基の置換、欠失、挿入といった文字解析・判断処理が合計で最低で k 回必要であるものとする。値 k は、サブシーケンスSと相補配列Qとの間の最大編集距離として参照することができる。ここで、プローブPに対する相補配列Qの塩基鎖長を m として定義すれば、相似度 r は、 $r = (m - k) / m$ として定義することができる。したがって相似度が r 以上とは、 r についての上記式を参照して、最大編集距離が $m(1 - r)$ 以下のサブシーケンスSと相補配列Qとを指

定することに対応する。

【 0 0 2 8 】

図 1 (b) は、上述した編集距離の決定する処理を具体的な実施の形態を使用して説明した図である。図 1 (b) には、サブシーケンス $S = \text{ATGCGCGTAAAGT}$ であって、相補配列 $Q = \text{ATGCGTTAACT}$ である場合の実施の形態をもって説明されている。図 1 (b) に示した実施の形態では、サブシーケンス S と、相補配列 Q との間の編集距離は 4 となり、相似度 r は、 $r = (11 - 4) / 11 = 0.6364$ となる。図 1 (b) には、部分配列 T_p と相補配列 Q とをそれぞれ整列（以下、上述した処理を、アラインメントとして参照する。）させて示している。図 1 (b) に示した実施の形態では、欠失が 2 回あり、置換および挿入がそれぞれ 1 回なので、編集距離は 4 である。

【 0 0 2 9 】

本発明においては、所定のターゲット塩基配列 T において、ターゲット塩基配列 T と、相補配列 Q との間には、ターゲット塩基配列 T に相補配列 Q の $(m - k)$ 個以上の塩基が順番を違えずに出現することがない場合、ターゲット塩基配列 T には相補配列 Q との編集距離が k 以下のものは存在しないといえることができる。プローブ P と相補配列 Q との間には、図 1 (a) に示した関係が存在するため、プローブ P から生成された相補配列 Q とターゲット塩基配列 T についての情報を得ることができれば、プローブ P とサブシーケンス S とについての関係を容易に得ることが可能となる。

【 0 0 3 0 】

一方で、本発明においては、所与のプローブ P が結合してしまう可能性のあるターゲット塩基配列をスクリーニングすることもできる。図 2 は、本発明において部分配列 T_p と相補配列 Q との関係において、相補配列 Q の特異性を判断する場合の処理を示した図である。図 2 では、複数のターゲット塩基配列がスクリーニングの対象とされており、プローブ P がターゲット塩基配列 T_1 のみに特異的に結合するか否かを判断する実施の形態の概略図である。ターゲット塩基配列 T_1 には、プローブ P の相補配列 Q がそのまま含まれているのが示されており、また T_a についても、相補配列 Q_a が、ちょうど $(m - a)$ の編集距離 (a は、 k よりも

大きい正の正数である。) で存在しているのが示されている。また、ターゲット配列 T_b には、 $(m-b)$ (b は、 k よりも小さい正の整数である。) の編集距離部分配列が存在しており、この結果、プローブは、ターゲット塩基配列 T_1 および T_b について結合する可能性があることが示される。この場合、本発明によれば、結合可能性の情報と共にサブシーケンスとの相似度および塩基鎖長がユーザに提示されるので、プローブ P をさらに特性を付与するべく設計することも可能である。

【0031】

B. 本発明のスクリーニング方法の処理

上述した特性を効率的に利用することにより、本発明では、ターゲット塩基配列が、プローブ P として適切か否か、また他のプローブ P を提供する必要があるか否かの判断を、高精度な逐次的解析するよりも、より高速に確認することが可能となる。すなわち、本発明によれば、ターゲット塩基配列 T にプローブ P がハイブリダイズしないことの保証が可能になる。図3には、本発明の方法の本質的な処理のフローチャートを示す。図3に示すように、本発明の塩基配列のスクリーニング方法は、ステップ $S10$ においてプローブ P とすべき塩基配列およびターゲット塩基配列 T を受け取る。この際の入力としては、シーケンサなどによって実際に得られた塩基配列でも良い。さらに、プローブ P は、塩基配列のシンセサイザなどによって合成することもできる。また、ターゲット塩基配列については、例えばDNAまたはRNAといった塩基配列を格納したデータベースから、インタラクティブにユーザがターゲット塩基配列を選択することもできるし、ほ乳類、真核微生物などのlive tissueから採取したDNAサンプルを使用することもできる。図3における本発明のスクリーニング方法は、次いで、ステップ $S12$ へと進み、プローブ P から相補配列 Q を生成し、適切な記憶領域に格納すると共に、最大編集距離 k を受け取り、適切な記憶領域に格納する。

【0032】

ステップ $S14$ では、検査すべきターゲット塩基配列 T について、関数 `largest_prefix_with_common_subsequence()` を適用して、その時点で検討を加えているターゲット塩基配列 T について、相補配列 Q がターゲット塩基配列 T のサブシ

一ケンスとして存在するか否かを判断することにより、結合性の評価を行なう。さらの本発明の処理は、ステップS16へと進み、評価結果をコンピュータにおける適切な記憶領域に格納する。さらにステップS16では、格納された評価結果を読み出して「結合する」、「結合しない」の判断を実行し、その結果がユーザに返され、ユーザに対して判断の結果が提供される。

【0033】

図4は、本発明のスクリーニング方法の第2の実施の形態の処理を示したフローチャートである。図4に示した実施の形態では、対象とされるターゲット塩基配列Tは、複数 $T_1 \sim T_w$ として存在し、ターゲット塩基配列ごとに評価結果が作成され、記憶領域に格納される。図4に示した実施の形態ではステップS20において、プローブPとするべき塩基配列および複数のターゲット塩基配列 $T_1 \sim T_w$ を受け取り、それぞれの格納部に格納する。ステップS22において、プローブQから相補配列Qを決定し、さらに編集距離k、評価を繰り返して実行する部分配列の長さの単位(length)と、重なり合いの長さ(overlap_length)といった変数の入力を受け取る。ステップS24においてカウンタtを、 $t=1$ に設定し、ステップS26においてターゲット塩基配列に対して結合性の評価を実行する。ステップS26においてターゲット塩基配列について評価を行い、その結果を記憶部に格納する。ステップS28では、評価結果にアクセスして結合可能なサブシーケンスの有無などの情報を生成する。

【0034】

ステップS30においてカウンタtをインクリメントし、ステップS32において、tがwよりも小さいか否かを判断する。ステップS32の判断において $t \leq w$ の場合(yes)には、ステップS34に進み、 T_{t+1} のターゲット塩基配列を記憶領域から読み出し、ステップS26へと戻り、 T_{t+1} のターゲット塩基配列についての評価結果を作成して格納する。ステップS32の判断が $t > w$ 場合(no)には、すでに評価を行うターゲット塩基配列について評価を終了しているので、ステップS36へと進み、評価結果を判断してユーザに結合性の判断結果を提示する。

【0035】

図 5 は、本発明において結合性を評価する処理を実行する関数である、`largest_prefix_with_common_subsequence()` のフローチャートを示した図である。図 5 に示した本発明における結合性の評価の処理は、ステップ S 4 0 の初期化処理から開始し、`positions[i]` の値および編集距離カウンタ `i` の値を初期化する。ここで、`positions[i]` は、関数 `next_char_position()` で与えられる値を返すか、または関数 `next_char_position()` で与えられる値と `positions[i-1]` で与えられる値のうちの小さい方の値を返す関数である。また、関数 `next_char_position()` は、引数としてターゲット塩基配列、編集距離カウンタ `i`、相補配列に存在する塩基の文字 `c` とを引数として使用し、`next_char_position(T, i, c)` は、ターゲット塩基配列 `T` の塩基鎖の先頭から数えて、`T[i..n]` の中で最初に出てくる文字 `c` の位置を返す関数である。また、文字 `c` の塩基が `T[i..n]` の中に出現しない場合には、値 `(n+1)` を返す関数である。

【 0 0 3 6 】

ステップ S 4 2 においては、評価を行う相補塩基の塩基位置を示す変数である `probe_position` の値および `min_k` の値を初期化する。この `min_k` の値は、`min_k` よりも小さな編集距離はありえない値を規定するものであり、部分配列と相補配列 `Q` との間の塩基の不致の最小値を与える尺度である。すなわち、`min_k=0` は、サブシーケンス `S` と、相補配列 `Q` とが完全に一致することに対応する。同時に `min_k` は、停止判断を行うための停止条件値としても使用され、さらにプローブ `P` とサブシーケンス `S` との間の塩基誤差の尺度を与える。ステップ S 4 4 においては、相補塩基 `Q` に含まれる塩基数に対応する変数 `probe_position` が相補塩基 `Q` の塩基数である `m` よりも大きいか否かを判断する。ステップ S 4 4 における判断において変数 `probe_position` が `m` よりも大きくなった場合 (yes) には、すでに評価すべき相補塩基の塩基は残されていないので、ステップ S 4 8 へと進む。この段階では、編集距離カウンタ `i` は `min_k` よりも大きいので、`m-k` 以上の鎖長のサブシーケンスが存在する可能性がまだ残されているので、`positions[k]` の値を出力し、記憶部に格納する。

【 0 0 3 7 】

一方、ステップ S 4 4 の判断において変数 `probe_position` が `m` 以下の場合 (no)

には、ステップ S 5 2 に示す評価処理を実行する。評価処理は、より詳細には後述するが、`position[i]` の値を決定し、評価結果を記憶部に格納する処理を実行させる。詳細には後述するように評価処理が所定の停止条件を満たすことにより終了すると、処理はステップ S 4 6 において `min_k` の値が最大編集距離 `k` よりも大きくなったか否かを判断する。`min_k` が `k` よりも大きい場合 (yes) には、 $(m-k)$ 以上の共通サブシーケンスが存在しないので、ステップ S 5 0 に進んで、 $(n+1)$ を `positions[i]` の値として出力させ、評価処理を停止させる。ステップ S 4 6 の判断における “true” の値が評価停止を行うための信号として生成される。また、ステップ S 4 6 の判断において、`min_k` の値が `k` より大きくない場合 (no) には、その後処理は、ステップ S 5 4 へと進み、`probe_position` の値をインクリメントして、処理をステップ S 4 4 へと戻し、相補配列 Q における次の塩基の評価を再度実行させる。ステップ S 4 4 からステップ S 5 2 の処理は、`min_k` の値を動的に変化させることで評価処理の停止条件を動的に変更させながら、ステップ S 4 6 の判断が肯定的な結果 (yes) を返すまで繰り返される。

【 0 0 3 8 】

図 6 は、図 5 のステップ S 4 0 に示した初期化処理の処理を示した図である。図 6 において示した初期化処理は、ステップ S 6 0 において編集距離カウンタ `i` を 0 に初期化し、ステップ S 6 2 およびステップ S 6 4 において、編集距離カウンタ `i` が最大編集距離 `k` よりも大きくなるまで `position[i]=0` としてセットする。同時に編集距離カウンタ `i` を、 $i=i+1$ へとインクリメントする。この初期化処理は、編集距離 `k` 以下に対し、`positions[i]` の値が確実に 0 となるように設定するために使用される。

【 0 0 3 9 】

図 7 は、図 5 において説明したステップ S 5 2 の評価処理の詳細なプロセスを示したフローチャートである。図 7 に示した評価処理は、ステップ S 7 0 において `max_k` を最大編集距離の値 `k`、または `probe_position` の値のうち、小さい方の値として設定する。ステップ S 7 2 では、編集距離カウンタ `i` が停止条件である `min_k` よりも小さいか否かを判断する。編集距離カウンタ `i` の値が `min_k` 以上の場合には (no) には、ステップ S 7 4 に進んで編集距離カウンタ `i` が 0 か否かを判断する

。また、ステップ S 7 2 の判断において編集距離カウンタ i が min_k より小さいと判断される場合 (yes) には、評価処理を停止し、図 5 のステップ S 5 4 へと分岐する。

【0040】

ステップ S 7 4 の判断において、編集距離カウンタ i が 0 の場合 (yes) には、ステップ S 7 6 へと進んで $\text{positions}[i]$ の値を $\text{next_char_position}(T, \text{positions}[i]+1, Q(\text{probe_position}))$ として与える。

【0041】

一方、ステップ S 7 4 の判断において、 $i=0$ ではないと判断された場合 (no) には、処理は、ステップ S 7 8 へと分岐し、 $\text{positions}[i]$ の値を、 $\text{next_char_position}$

$(T, \text{positions}[i]+1, Q(\text{probe_position}))$ の値、または $\text{positions}[i-1]$ の値のいずれか小さい方の値として設定する。したがって、 $\text{positions}[i]$ がその時点で評価している塩基数 n 以下の値を返す場合には、所定の最大編集距離 k 内において一致する塩基配列が存在する可能性があることを意味する。ステップ S 7 6 およびステップ S 7 8 において算出された $\text{positions}[i]$ の値は、評価結果として適切な記憶部に格納される。

【0042】

その後、処理はステップ S 8 0 へと進んで、 $\text{positions}[i]$ が n よりも大きくなったか否かを判断する。上述したように、関数 $\text{next_char_position}()$ は、評価している塩基に該当するものが見出されない場合であってもその値として相補配列 Q の塩基数である n 以下の値を返す。したがって、ステップ S 8 0 の判断において、 $\text{positions}[i]$ の値が m 以下である場合 (no) には、さらに適合する可能性を判断するために、ステップ S 8 4 において編集距離カウンタ i を、 $i=i-1$ として降順に設定し、ステップ S 7 2 へと処理を戻してさらに判断を続行させる。また、ステップ S 8 0 の判断が (yes) の場合には、図 5 に示した $\text{largest_prefix_with_common_}$

$\text{subsequence}()$ により $\text{positions}[i]$ の値が $n+1$ として与えられ、図 7 の評価処理を終了させるべく、ステップ S 8 2 において $\text{min_k}=i+1$ として設定した後、ステップ

S 8 4 へと処理を戻す。ステップ S 8 4 では、 $i=i-1$ に設定した後、処理をステップ S 7 2 へと戻すので、ステップ S 7 2 において確実に判断結果として真の値 (yes) を返し、図 5 に示したステップ S 5 4 へと分岐させ、処理を繰り返す。

【 0 0 4 3 】

図 8 には、本発明のスクリーニング方法において、ターゲット塩基配列の評価を部分配列ごとに実行させる実施の形態を示す。図 8 に示した実施の形態では、ターゲット塩基配列 T に対して部分配列の所定の長さを規定する変数 `length` と、部分配列のターゲット塩基配列における位置を指定する `position` と、部分配列間の末端部における重なり合いを与える変数である `overlap_length` とを定義する。図 8 に示す本発明のスクリーニング方法の処理は、ターゲット塩基配列の塩基位置を、変数 `position`、変数 `length`、変数 `overlap_length` を使用して部分配列ごとに評価をすることを除き、図 5 ～図 7 に示した処理を適用することができるので、本発明の部分配列ごとに区切って評価を行う処理の本質的な部分のみを説明する。ステップ S 9 0 において、まず部分配列を指定する変数 `position` を初期設定する。次いでステップ S 9 2 において変数 `position` が n 以下であるか否かを判断する。ステップ S 9 2 の判断では、ターゲット塩基配列の最末端である n まで評価が実行されるまでは `position` は n 以下なので、ステップ S 9 2 の判断は、真 (yes) とされ、ステップ S 9 4 へと進み変数 `length` を定義する。この変数 `length` は、あらかじめ実験条件などを考慮して設定またはユーザ入力により定めることができるし、本発明においては、他の適切な関数、例えば図 5 に示した `largest_prefix_length_with_common_subsequence()` の戻り値を使用して、ステップ S 9 4 に示すようにして定めることもできる。

【 0 0 4 4 】

関数 `largest_prefix_length_with_common_subsequence()` は、上述したようにランダムで十分に長いターゲット塩基配列と、ランダムな相補配列 Q とが与えられた場合、 $4m$ の値を返す関数である。本発明においては上述した特定の関数を使用するのではなく、適切な値を取るように `length` を設定することができる。また、図 8 に示した実施の形態においても編集距離カウンタ i については、指定され

た部分配列ごとに初期化設定を行い、ターゲット塩基配列Tとして部分配列T_wを使用することにより、図5～図7に示した処理を繰り返して実行させることができる。ステップS96では、変数lengthの値が、塩基相互の距離にかかわらず結合する可能性がある程度高いことが想定される塩基鎖長であるcheck_lengthよりも短いかなかを判断する。

【0045】

ステップS96の判断において、lengthがcheck_lengthよりも短い場合(yes)には、ステップS104へと進んでSmith-Waterman法(P.D. Smith, M.S. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147: 195-197, 1981.)などの高精度な手法を使用して結合性の判断を行ない、check_lengthを加えて新たなposition値を生成して、ステップS92へと処理を戻すこともできる。この際、check_lengthの値をm/r以上の値とすることもできる。また、4mの値として設定すると、部分配列すべてについて「結合する可能性がある」と判断し、ステップS104の評価を高頻度に行うことになるので、本発明の高速性が阻害されてしまうことになる。このため、本発明においてはcheck_

lengthの値をm/r～4mの間の値として設定することができ、check_lengthの値として概ね2m程度の値を採用することが、安定した結果を与えることが多いことが見出された。

【0046】

一方、ステップS96の判断においてlengthがcheck_length以上の場合(no)には、ステップS98においてlengthと残りの塩基鎖長とを比較し、lengthが残りの塩基鎖長よりも短い場合(yes)には、さらにステップS102へと進んで、それまで評価を行った塩基位置にlengthを加え、さらに適切な重なり合いを与えることができるようにoverlap_length分を考慮して次の部分配列を指定する。図8に示した実施の形態では、その後再度ステップS92へと処理を戻し、ステップS92～ステップS98までの処理を繰り返す。また、lengthが残りの塩基鎖長を越えた場合(no)には、ステップS100においてposition=n+1に設定し、処理をステップS92へと戻す。この場合、ステップS92の判断において偽の値が

返され(no)、ステップS106において結合しないという評価結果が与えられる。

【0047】

図9には、図5～図7に示した処理を実行させるための擬似コードを示す。また、図10には、図8に示した部分配列ごとに評価を実行するための処理の擬似コードを示す。図10において使用された関数check_exactly(T, Q, k)は、ステップS102において使用される高精度の判定手法を示す。関数check_exactly(T, Q, k)は、ターゲット塩基配列Tと相補配列Qとが実際に結合するかどうかの判定を、従来の時間はかかるものの、正確に計算できるSmith-Watermanなどの方法を使用して計算し、結合しないならばtrueを、結合するならばfalseを返す。本発明においては、check_exactly(T, Q, k)としては、Smith-Watermanの方法を具体的には挙げることができるが、Smith-Watermanの方法以外にでも、いかなる方法を使用しても構成することができる。また、関数check_exactly(T, Q, k)について、結合しないことを高精度にユーザに対して提示するために、この関数を常にFALSEを返す関数にすることもできる。この場合には、「結合しない」という結果が出た場合は必ず結合しない、ということが言えるので、ハイブリダイゼーションの判定にも有用である。

【0048】

図11は、本発明のスクリーニング方法により与えられる評価結果をテーブルの形式で示した図である。図11に示した実施の形態においては、ターゲット塩基配列TがCGCGCATGAAであり、相補配列QがGCCCCATGCであって、編集距離k=3の場合に得られる実施の形態である。以下、図11を使用して本発明のスクリーニング方法により得られる評価結果について説明する。図11に示した表の縦欄は、編集距離カウンタiの値に対応し、横欄は、相補配列中位置を表すprobe_positionの値を示し、表中の値は、図7における評価処理の終了時のposition[i]の値を示している。図11に示した実施の形態では、最大編集距離k=3なので、編集距離カウンタiが3以下のものについてのみ計算が行われている。図11に示した実施の形態では、positions[i, probe_position]の計算は、最終的に図11に示した擬似コードの②まで計算が進み、positions[3]の値として7が得られてい

るため、返り値として7を返し、このときmin_kの最終的な値は、2となる。上述したpositions[i, probe_posiiton]は、positions[i, probe_posiiton]が、jの時のpositions[i]の値を、probe_position[i, j]とする。この配列データは、テーブルとして格納しておくこともできるし、本発明において説明するように、評価を行っているprobe_positionの変数に対するposition[i]の値のみを格納させておいても良い。

【0049】

図12(a)、(b)には、図11に示したpositions[0]およびpositions[1]についての値を返す処理を図式的に示す。図12(a)に示されるように、positions[0]の場合には、相補配列Qにおいて出現する塩基が単純にターゲット塩基配列において出現する位置となり、2, 3, 5で与えられる値がpositions[1, m]の値として与えられている

【0050】

一方でpositions[1]についてみれば、 $i \geq 2$ については、next_char_positionと、positions[i-1]の値のうちの小さい方の値が返り値とされるので、positions[0, 1]の値は2であるものの、相補配列Qのシトシン(C)がターゲット塩基配列の1番目に出現しているので小さい方の1が得られている。同様に相補配列Qの2番目のシトシン(C)は、ターゲット塩基配列Tの4番目に出現しており、next_char_positionの返り値は4となるものの、これは対応するpositions[0, 2]の値が3のため、小さい方の値3が返されている。

【0051】

さらに、相補配列Qの4番目の塩基であるシトシン(C)は、5番目に出現し、アデニン(A)は、6番目に出現し、チミン(T)は、7番目に出現し、7番目のグアニン(G)は、8番目に出現しており、これらの番号がpositions[1]の行を構成し、1, 3, 5, 6, 7, 8の行が与えられることになる。同様にして、position[2]については、相補配列Qの3番目のシトシン(C)がターゲット塩基配列Tの1番目に出現し1の値を与え、4番目のシトシン(C)が3番目に出現し、5番目のアデニン(A)が6番目に出現しているが、対応するpositions[1]のm=4の値である5が採用され、6番目のチミン(T)が7番目に出現しているも

のpositions[1]のm=5の値である6が採用されるので図11に示されるように、0, 1, 3, 5, 6, 7, 8の列が与えられているのが示される。

【0052】

図11に示された実施の形態であるk=3の場合、positions[1]の値がprobe_position=7の地点で8の値をとる。そのため、next_char_position (T, 9, "C" の値がn+1=11となり、nよりも大きな値なので、最終的なprobe_position=8の処理の際には、min_kの値が1だけインクリメントされ、2が与えられる。このmin_k=2の値は、ターゲット塩基配列および相補配列に共通するサブシーケンスの長さがm-2=6である、ということを示している。また、同じターゲット塩基配列と相補配列とを使用し、最大編集距離k=1として同様の計算を実行させると、編集距離カウンタi=0,1についてのpositions[1]およびpositions[0]の計算が実行され、評価結果が得られ処理を終了する。この場合、擬似コードの①まで計算が進み、n+1=11という値を返す。これは、m-1=7以上の長さの共通するサブシーケンスが存在しないことを示す。本発明においては、図11に示した評価結果テーブルとして必ずしも評価結果を蓄積する必要はなく、上述したように最大のpositions[i]の値についてだけ格納させるデータ形式で記憶部に評価結果を格納しておくこともできる。

【0053】

C. 本発明の塩基配列スクリーニング・システム

図13は、本発明のスクリーニング方法を実装した塩基配列スクリーニング・システムの該略的な機能ブロック図である。図13に示された塩基配列スクリーニング・システム（以下、システムと略する。）10は、概ね中央処理装置（CPU）および必要な記憶装置とを含むコンピュータを含んで構成されており、ターゲット塩基配列や相補配列の入力を受けて、スクリーニングのための評価処理を実行させている。

【0054】

さらに詳細に説明すると、本発明のシステム10は、相補配列格納部14と、ターゲット塩基配列格納部16と、最大編集距離格納部18と、評価処理部20とを含んで構成されている。相補配列格納部14、ターゲット塩基配列格納部1

6 および最大編集距離格納部 1 8 は、それぞれ適切なメモリまたはメモリ領域として構成されている。図 1 3 においては、機能ブロックを明確に表示させるべく、異なるブロックとして記述されているものの、本発明においては、実装上の都合により、単一のランダム・アクセス・メモリなどに割り当てられたメモリ領域として構成することもできる。相補配列格納部 1 4 は、入力部 1 2 からのプローブ P の入力を受け取り、相補性の関係を使用して生成された相補配列 Q を格納する。また、ターゲット塩基配列格納部 1 6 および最大編集距離格納部は、入力されたターゲット塩基配列のデータおよび最大編集距離 k を、それぞれ格納している。上述した入力部 1 2 は、本発明においては、キーボード、データ読み取り装置などを含んで構成することができ、データ読み取り装置は、例えばフロッピー（登録商標）・ディスク、ハードディスク、CD-ROM、DVD といった入出力媒体などに記憶されたプローブ P の塩基配列を読み出し、本発明のシステムへとプローブ P のデータを送っている。

【 0 0 5 5 】

評価処理部 2 0 は、本発明のスクリーニング方法を実行させるためのプログラムを、図示しないハードディスクなどから CPU が読み出して、ソフトウェア的に構成される機能ブロックとして実装される。評価処理部 2 0 は、図 5 に示した `largest_prefix_with_common_subsequence` により構成され、動的に評価を停止させる停止判断部 2 0 a と、評価実行部 2 0 b とを含んで構成されている。評価処理部 2 0 は、相補配列格納部 1 4、ターゲット塩基配列格納部 1 6、および最大編集距離格納部 1 6 から、それぞれ相補配列 Q、ターゲット塩基配列 T、最大編集距離 k を読み出し、本発明のスクリーニング方法を適用して、共通するサブシーケンスの評価・判断を実行する。停止判断部 2 0 a は、`min_k` の値と最大編集距離 k とを比較して、`min_k` の値が k よりも大きくなった場合には、ユーザの要求する範囲内において「結合する」と判断されるシーケンスがないものと判断して、評価実行部 2 0 b に対してそれ以後の `probe_position` について評価を行わないように指令する。この指令を受け取った評価処理部 2 0 は、計算された評価結果を、本発明における特定の実施の形態では、`positions[i, probe_position]` のフォーマットとして記憶部 2 2 へと書き込みを行なう。本発明の特定の実施の形態

においては、評価結果は、図 1 1 に示した評価結果テーブルとして構成することが後のさらなる評価のための情報を与える点では好ましいが、本発明においてはまた、必要とされる `positions[i, probe_position]` の値だけを選択して格納することもできる。

【 0 0 5 6 】

評価処理部 2 0 は、評価結果テーブルの作成を終了すると、`positions[i]` の値を参照し、所定の `probe_position` の値に対応する `position[i]` の値をルックアップし、`probe_position` に対応するコラムにおいて `probe_position` 以上の返り値が含まれているか否かを判断する。`probe_position > m` で、かつ `min_k` の値が `k` よりも大きくなった場合には、要求された最大編集距離 `k` に対応する範囲内で共通するサブシーケンスが見出されないと判断し、「結合しない」という判断結果をユーザに対して表示部 2 4 において表示させる。また、`probe_position > m` であって、`min_k < k` である場合には、当該 `positions[i]` の編集距離カウンタ `i` において、`min_k = i + 1` であり、シーケンス長が `(m - min_k)` で与えられる共通するシーケンスがあるものと判断される。この判断結果は、表示部 2 4 を通してユーザに「結合する」画面表示と共に、シーケンスの塩基配列およびシーケンス長といった付帯的情報を表示させることにより、提供される。ユーザに対して提供された判断結果は、ユーザの必要性に応じて、プリンタからハードコピーされるか、またはフロッピー（登録商標）・ディスク、ハードディスク、CD-ROM、DVD、フラッシュメモリなどの記憶媒体に格納することができる。

【 0 0 5 7 】

図 1 4 は、本発明のスクリーニング方法の第 2 の実施の形態を実装したシステムの実施の形態を示した図である。図 1 4 に示したシステム 1 0 は、評価処理部 2 0 の構成を除き、概ね図 1 3 において説明したと同一の構成とされている。このため、図 1 4 に示したシステム 1 0 について、評価処理部 2 0 について詳細に説明する。図 1 4 に示した評価処理部 2 0 は、停止判断部 2 0 a と、評価実行部 2 0 b と、部分配列指定部 2 0 c とを含んで構成されている。部分配列指定部 2 0 c は、まず最初に評価を開始する部分配列を、指定データ格納部 2 8 から `position` の値および `length` の値などを読み出して決定する。指定された部分配列は、

評価実行部20bへと渡されて、図13において説明したと同様に評価が実行される。停止判断部20aは、評価実行部20bへと評価の停止を指令すると共に、部分配列指定部20cに対しても部分配列について評価が終了したことを指令し、さらに評価を行うべき部分配列がある場合には、次に評価する部分配列を指定して、評価実行部20bに部分配列のデータを転送する。

【0058】

図15は、本発明のシステム10の第2の実施の形態を示した図である。図14に示した本発明の第2の実施の形態のシステムは、ネットワーク30を介して、ターゲット塩基配列TおよびプローブPの塩基配列のデータをユーザから受信する。受信されたデータは、送受信部34の入力バッファに格納された後、相補配列格納部14、ターゲット塩基配列格納部16、最大編集距離格納部18、および指定データ格納部28に格納される。この場合、ユーザは直接ターゲット塩基配列Tを送信することもできるが、cDNAデータベースなど、システム10がターゲット塩基配列Tについての情報を格納したデータベースを含んで構成される場合には、評価処理部20は、ネットワークを介してユーザから受け取ったターゲット指定データを参照して、データベース32にアクセスして該当するターゲット塩基配列を取得し、取得したターゲット塩基配列Tをターゲット塩基配列格納部16に格納する。評価処理部20は、図14において説明したと同様の処理を実行して評価結果を生成し、記憶部22に格納すると共に、評価結果を送受信部34へと送り、ネットワーク30を介してユーザに評価結果を送出して、塩基配列データベースに格納された塩基配列をスクリーニングするためのプローブPの有効性についての情報をユーザに返す構成とされている。

【0059】

【実施例】

本発明のスクリーニング方法の効果を検証するため、Windows（登録商標）2000を搭載したパーソナル・コンピュータ（Intel社製のPentium（登録商標）4:2.2GHzプロセッサを実装）に本発明のプログラムを実装して、従来の高精度法（Smith-Waterman法）と対比して計算速度の評価検討を行った。使用したターゲット塩基配列は、ライブラリから入手した4,639,221bpの塩基鎖長の大腸菌のゲノム（Ge

n Bank Accession番号:NC_000913)を使用した。実験用のパーソナル・コンピュータには、部分配列ごとに評価を実行する図9および図10で示した擬似コードのアルゴリズムを実装した。なお、check_exactlyの関数としては、Smith-Waterman法を使用した。

【0060】

入力する相補配列Qは、大腸菌の上述したゲノムの一部を使用して塩基を削除することにより塩基鎖長が約580bp程度の塩基鎖を生成し、複数のプローブPとした。最小のウィンドウサイズ(check_length)をプローブの長さの約2倍に相当する1158と固定した。ハイブリダイゼーションすることができる塩基鎖の相似度を、編集距離kを設定して、相似度を0.80、0.85、0.90、0.95となるように設定し、相補配列の相似度の異なる4回の実験を行った。また、比較のため、従来技術のSmith-Waterman法の計算時間も計測した。その結果を図16に示す。

【0061】

図16に示された結果によれば、本発明の方法によるスクリーニングの計算速度は、約2倍～約20倍であることが確認された。計算速度は、相似度を高く設定できる場合は、最大編集距離kを小さく設定するので、プローブPの塩基配列についての反復計算を制限することができ（すなわち、編集距離カウンタの反復回数がk以下となる）、判断を行う塩基の絶対数が減少することから高速化が可能であることが示された。すなわち、本発明によれば、従来手法とは異なり、相似度を高くすればするほど高精度な計算が可能となることがわかる。

【0062】

図17には、図16に示された表の結果を、縦軸を計算時間（s）、横軸を相似度（%）としてプロットしたグラフ図である。図17に示されるように、相似度が高くなればなるほど判断結果を得るための時間が短縮することが示されている。また、相似度100%における判断結果を得るための時間は、比較例であるSmith-Waterman法の結果に比べて少なくとも数10倍改善されているのが示された。

【0063】

以上、本発明について、図面に示した具体的な実施の形態をもって説明してき

たが、本発明の上述した各機能を実現する手段または部分は、コンピュータ実行可能なプログラム言語により記述されたソフトウェアまたはソフトウェア・モジュール群として構成することができ、必ずしも図面に記載した機能ブロックとして構成される必要はない。また、本発明の塩基配列スクリーニング・システムでは、各機能要素は、必要に応じていかなる機能モジュールと共に構成することができ、図面に示された特定の実施の形態に限定されるものではない。

【 0 0 6 4 】

さらに、本発明のプログラムは、種々のプログラミング言語、例えばFORTRAN、C言語、C++言語、JAVA（登録商標）などを使用して記述することができ、本発明のプログラムを記述したコードは、磁気テープ、フレキシブル・ディスク、ハードディスク、コンパクト・ディスク（CD）、光磁気ディスク、デジタル・バーサタイル・ディスク（DVD）といったコンピュータ可読な記録媒体に保持させることができるし、伝送媒体としても頒布することができる。

【 0 0 6 5 】

本発明によれば、少なくとも相補性の低いプローブを選択的に排除することにより、従来の高精度なアラインメントによる手法よりも高速に健全性の高いプローブに関する情報をユーザに提供することが可能となる。また、本発明の方法をプローブの1次スクリーニングとして適用し、スクリーニング後に従来の高精度なアラインメントによる手法と組み合わせることで、プローブのスクリーニング処理を、より高速化することが可能な塩基配列をスクリーニングするためのシステム、そのための方法、該方法をコンピュータに対して実行させるためのプログラム、該プログラムを記憶したコンピュータ可読な記録媒体およびネットワークを介して塩基配列のスクリーニングを実行させるためのサーバを提供することができる。

【図面の簡単な説明】

【図 1】 本発明におけるターゲット塩基配列と、プローブと、相補配列との関係を示した図。

【図 2】 本発明におけるターゲット塩基配列のスクリーニングを行う実施の形態の概略図。

【図 3】 本発明のスクリーニング方法の第 1 の実施の形態の処理を示したフローチャート。

【図 4】 本発明のスクリーニング方法の第 2 の実施の形態の処理を示したフローチャート。

【図 5】 本発明において結合性の評価を行う処理の概略的なフローチャート。

【図 6】 本発明において結合性の評価を行う処理のための初期化処理を示した図。

【図 7】 本発明において結合性の評価のための評価処理の詳細を示したフローチャート。

【図 8】 本発明のスクリーニング方法の第 2 の実施の形態の詳細な処理を示したフローチャート。

【図 9】 本発明の結合性の評価を行うための処理を実行させるプログラムの擬似コード。

【図 1 0】 本発明の第 2 の実施の形態において部分配列を指定して結合性を評価するためのプログラムの擬似コード。

【図 1 1】 本発明における評価結果の出力データ構造の実施の形態を示した図。

【図 1 2】 本発明における評価結果を得るための図 9 に示した擬似コードの実行する処理を示した図。

【図 1 3】 本発明の塩基配列スクリーニング・システムの第 1 の実施の形態を示した図。

【図 1 4】 本発明の塩基配列スクリーニング・システムの第 2 の実施の形態を示した図。

【図 1 5】 本発明の塩基配列スクリーニング・システムの第 3 の実施の形態を示した図。

【図 1 6】 本発明の実施例の結果を比較例の結果と共に示した図。

【図 1 7】 本発明の実施例および比較例の結果をプロットして、評価結果を得るまでに必要な時間を、相似度 1 0 0 % において比較した図。

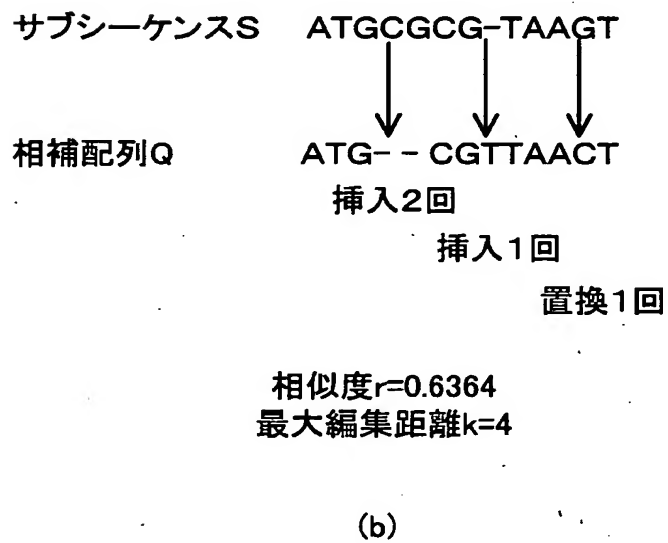
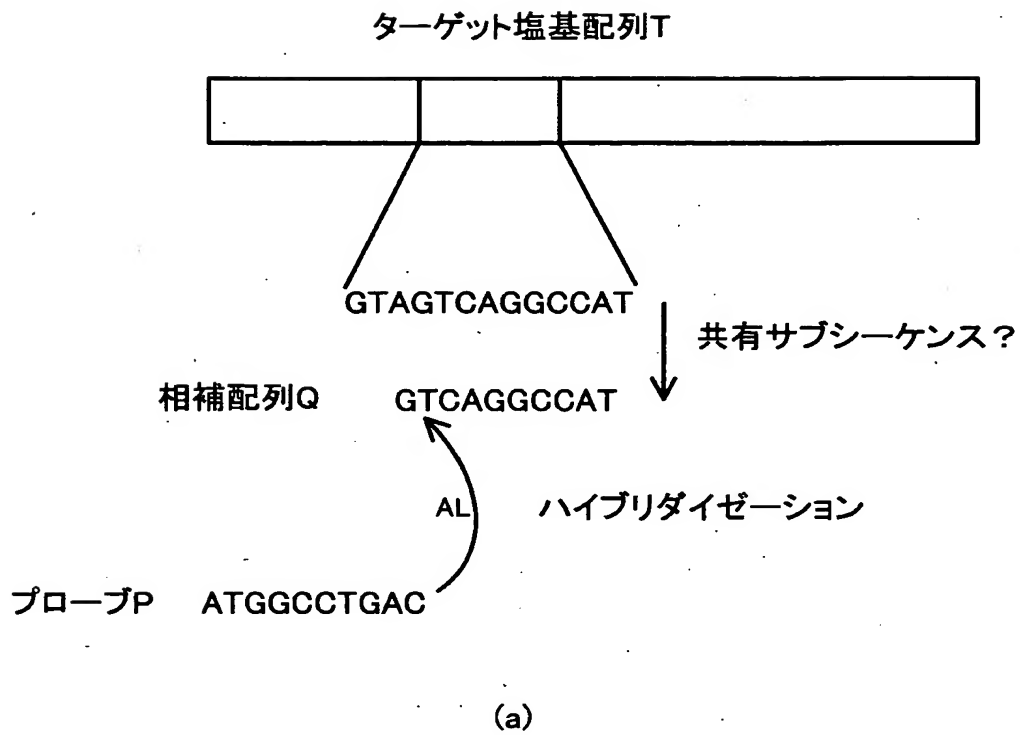
【図 1 8】 ターゲット塩基配列と、プローブとの関係を示した図。

【符号の説明】

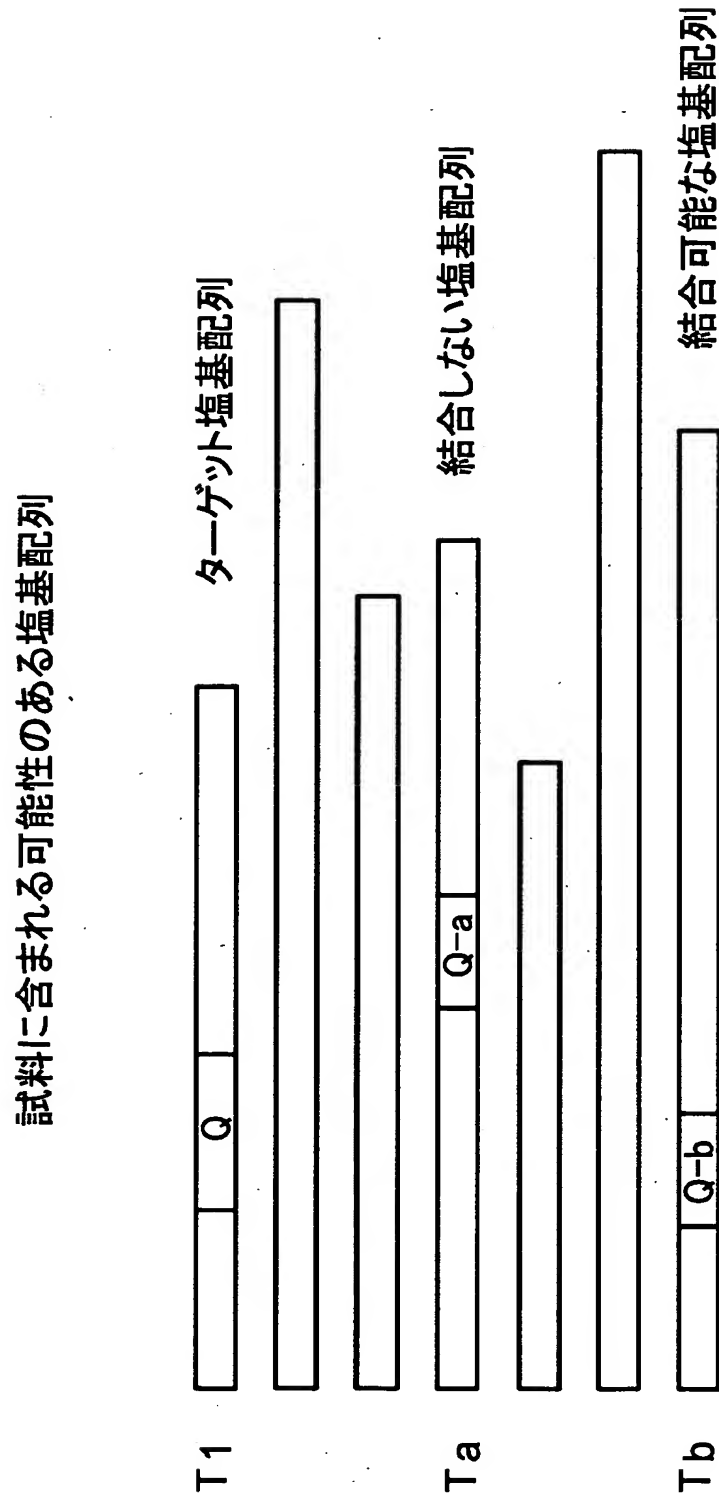
- 1 0 …塩基配列スクリーニング・システム
- 1 2 …入力部
- 1 4 …相補配列格納部
- 1 6 …ターゲット塩基配列格納部
- 1 8 …最大編集距離格納部
- 2 0 …評価処理部
- 2 2 …記憶部
- 2 4 …表示部
- 2 6 …出力部
- 2 8 …指定データ格納部
- 3 0 …ネットワーク
- 3 2 …データベース
- 3 4 …送受信部

【書類名】 図面

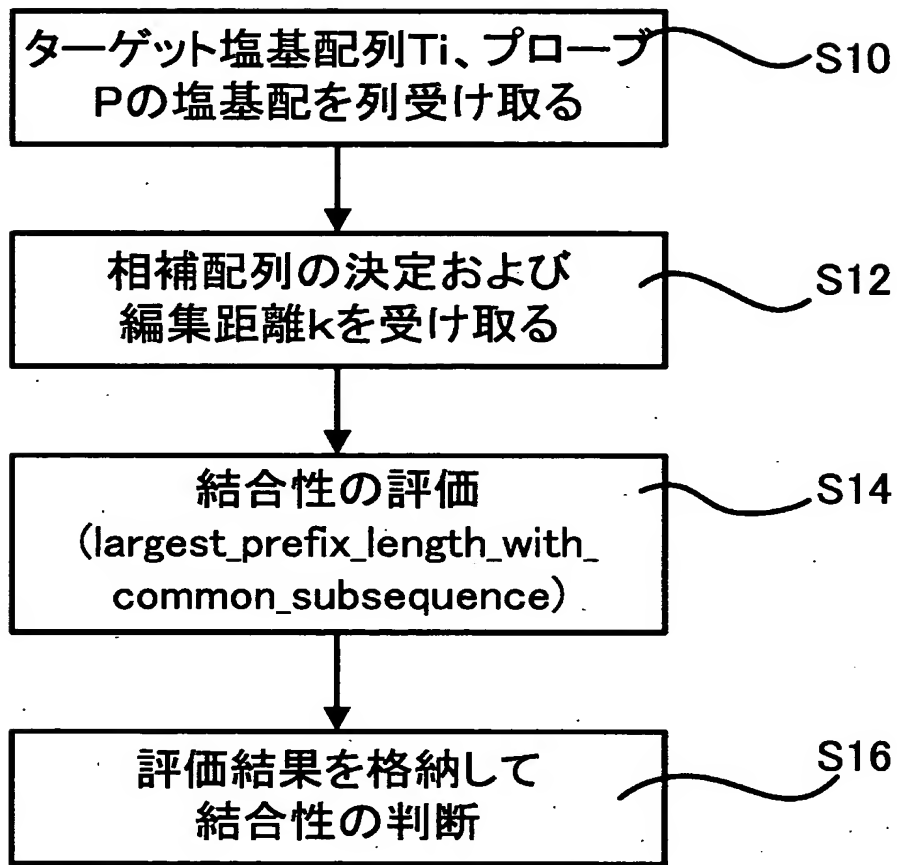
【図 1】



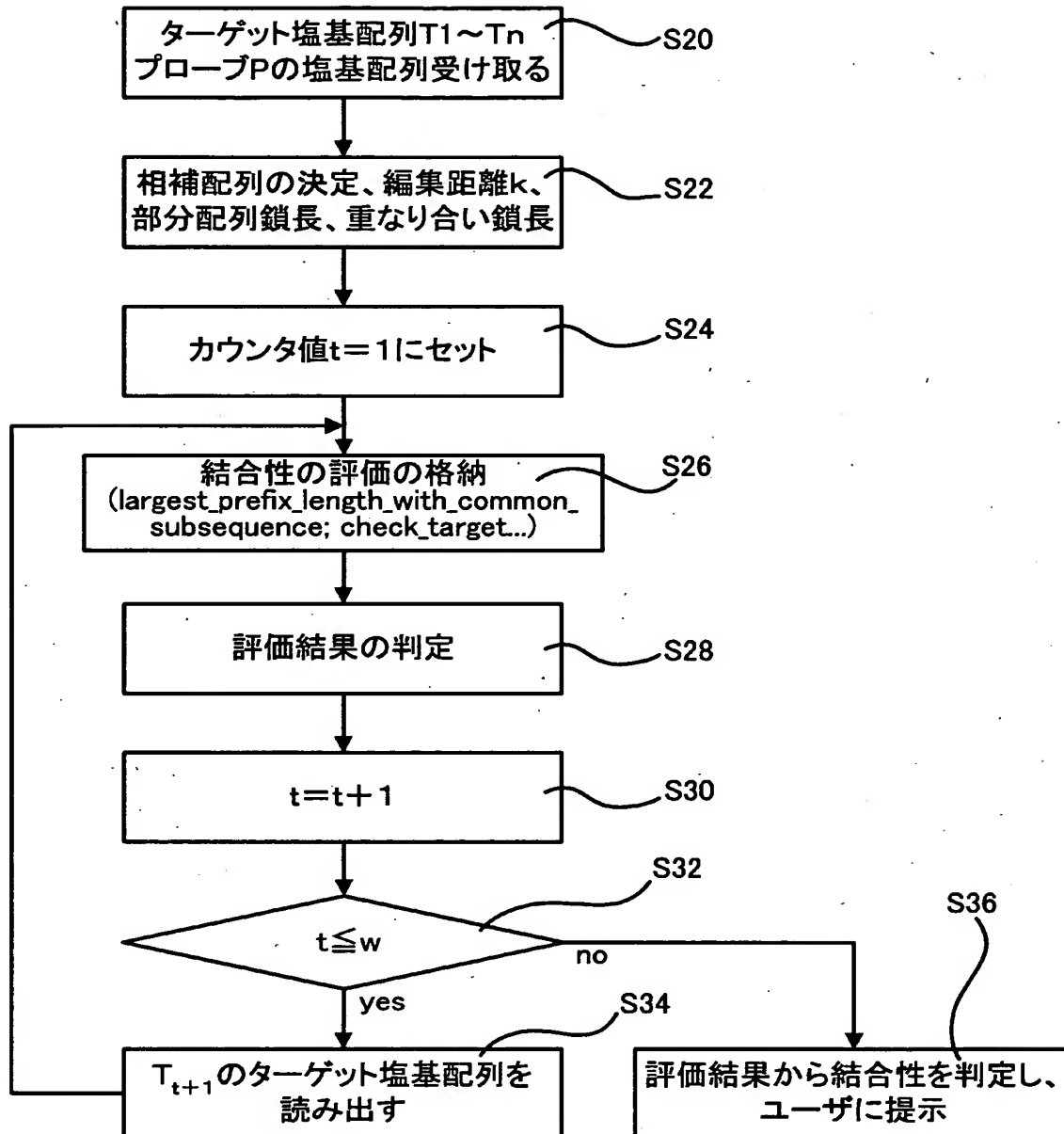
【図 2】



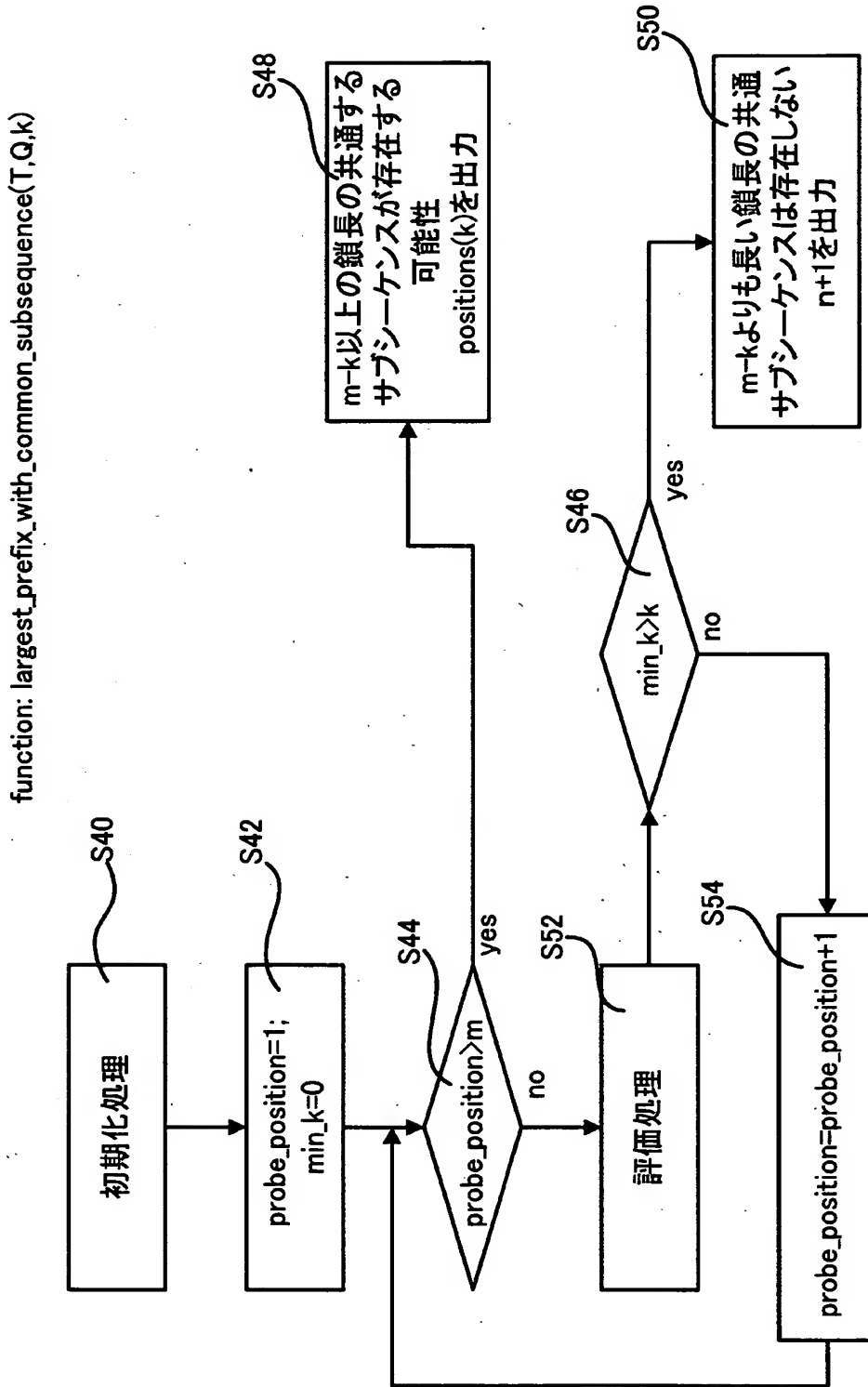
【図 3】



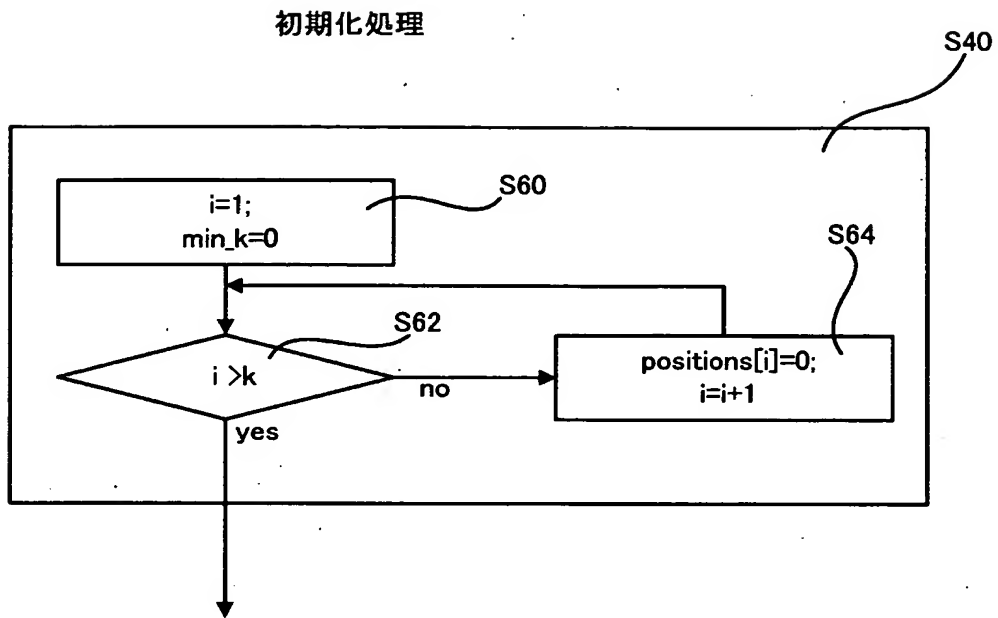
【図 4】



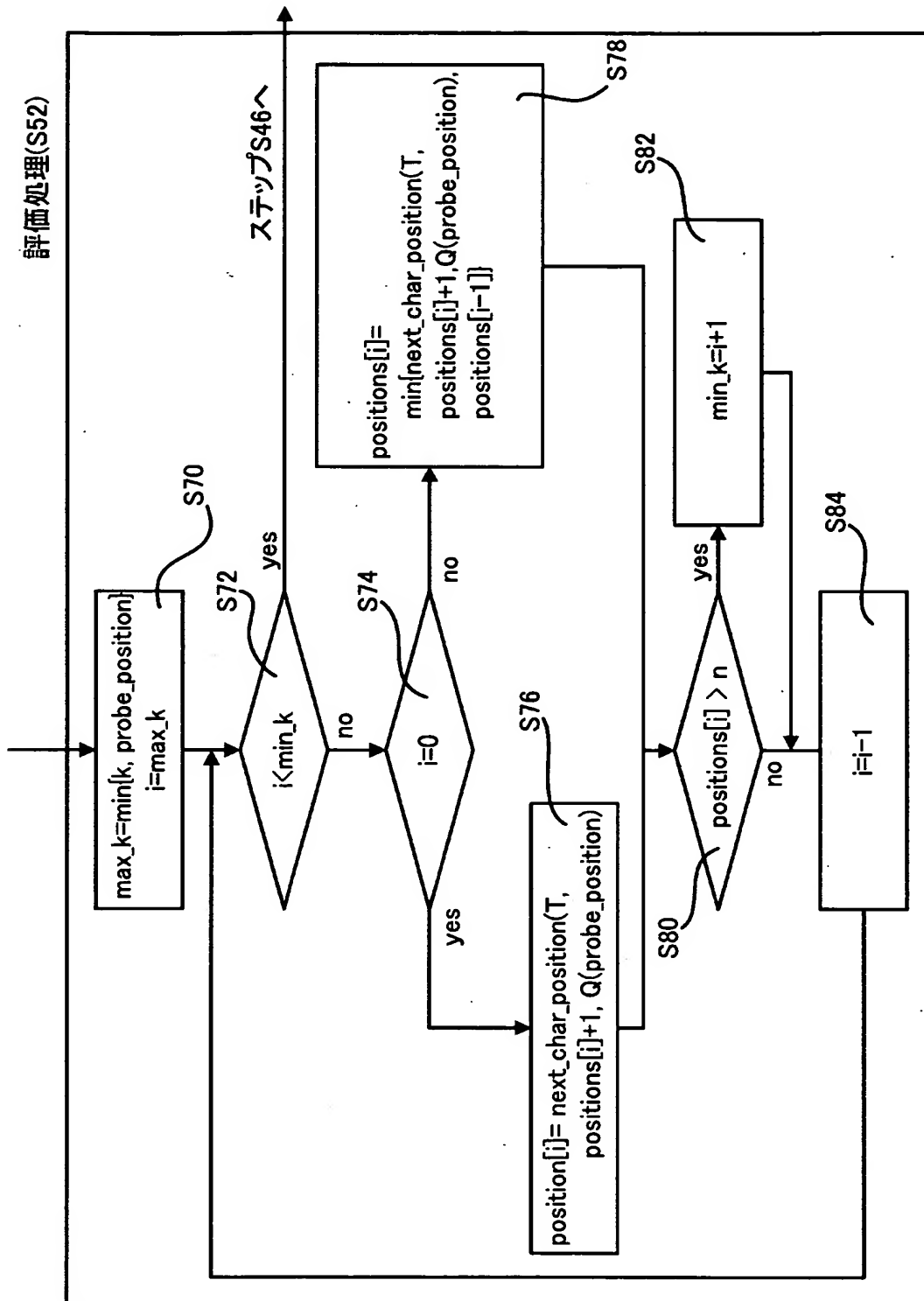
【図 5】



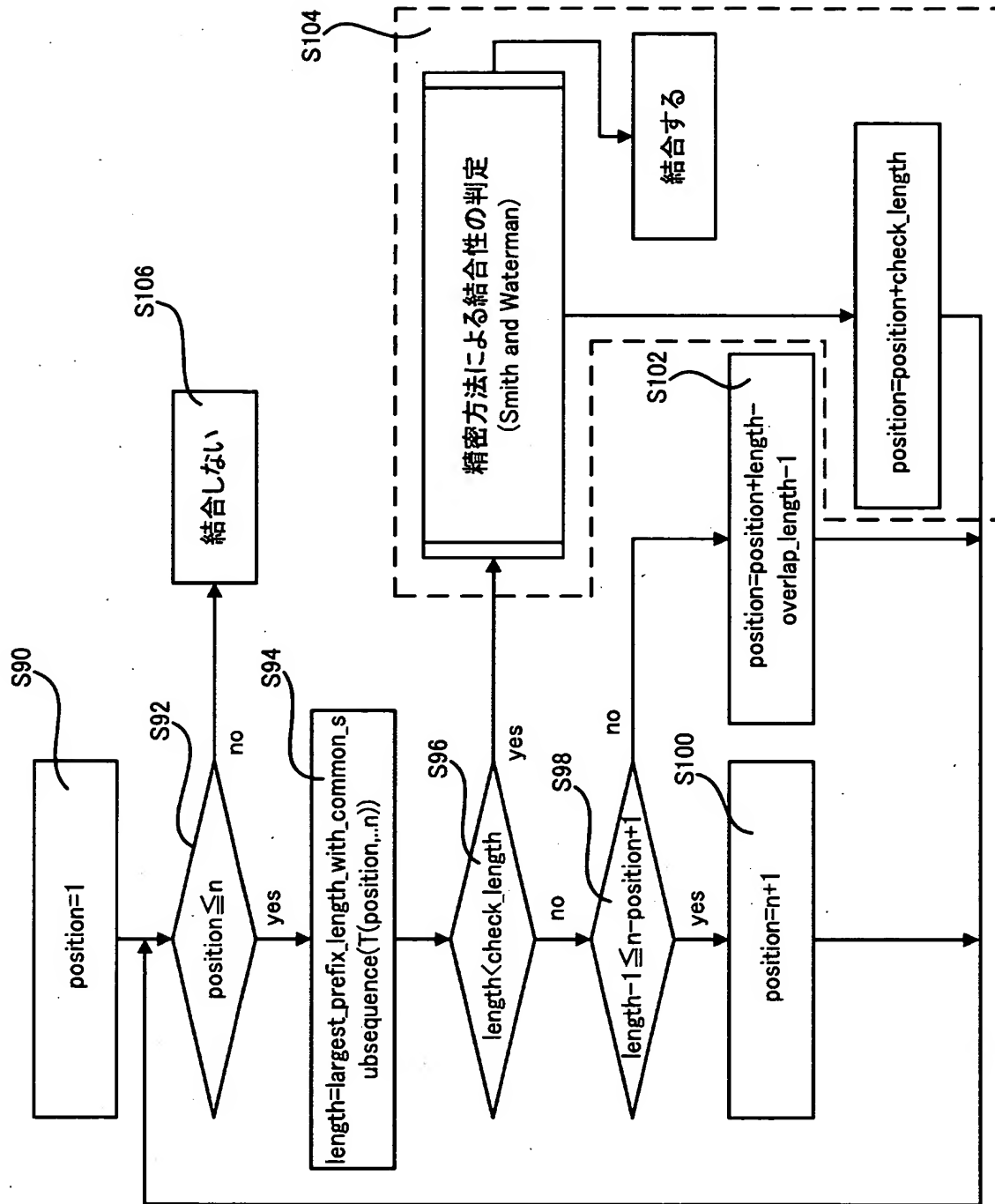
【图 6】



【図 7】



【図 8】



【図 9】

```

//関数 largest_prefix_with_common_subsequence

largest_prefix_with_common_subsequence(配列T[1..n], 配列Q[1..m], k) {
    for (i=0; i<=k; i++) positions[i] = 0;
    min_k = 0;
    for (probe_position=1; probe_position<=m; probe_position++) {
        max_k = min {k, probe_position};
        for (i=max_k; i>=min_k; i--) {
            if (i==0) {
                positions[i] = next_char_position(T, positions[i]+1,
Q[probe_position]);
            } else {
                positions[i] = min { next_char_position(T, positions[i]+1,
Q[probe_position]),
                                positions[i-1] }
            }
            if (positions[i] > n) {
                min_k = i+1;
            }
        }
        if (min_k > k) {
            ① m-k以上の長さの共通のサブシーケンスは存在しない
            return m+1;
        }
    }
    ② m-min_k以上の長さの共通のサブシーケンスが存在する
    return positions(k)
}

```

【図 1 0】

```

//関数 check_target

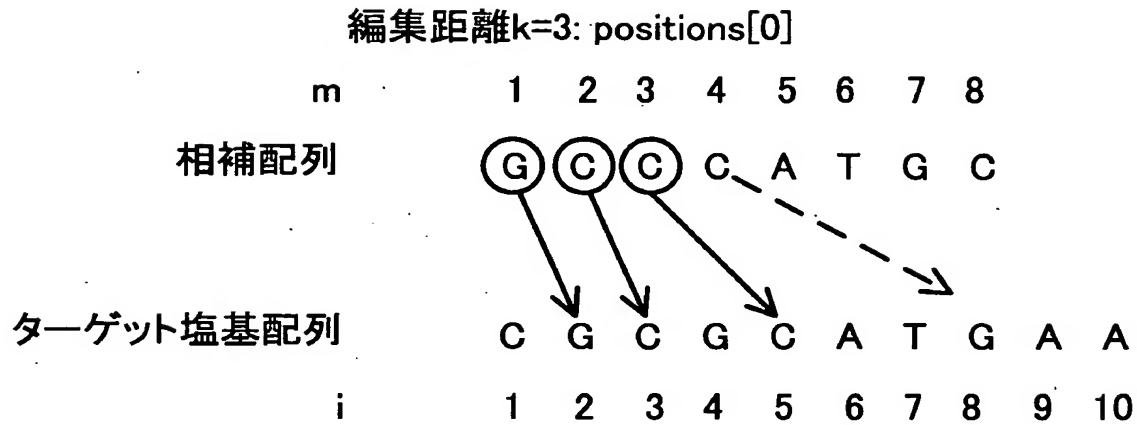
check_target(配列T[1..n], 配列Q[1..m],k, overlap_length, check_length){
    position=1;
    while(position <= n) {
        length = largest_prefix_length_with_common_subsequence(
                        T[position..n], Q[1..m], k);
        if (length < check_length) {
            if (check_exactly(
                    T[position..(min(n, position+check_length))], Q, k)) {
                position = position + check_length;
            } else {
                return ("TとPは結合する")
            }
        } else if (length - 1 <= n - position + 1){
            position = position + length - overlap_length - 1;
        } else {
            position = n+1;
        }
    }
    return ("TとPは結合しない");
}

```

【図 1 1】

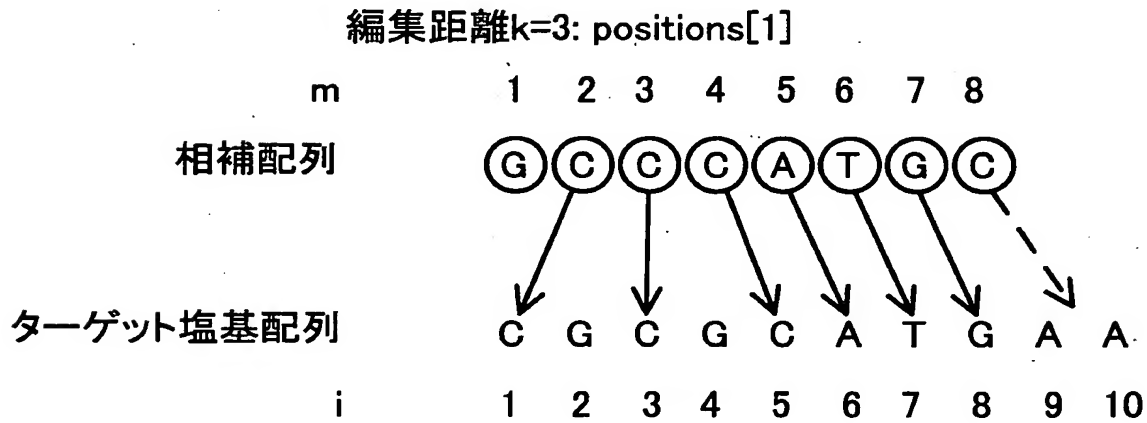
	1	2	3	4	5	6	7	8
positions[0]	2	3	5	-	-	-	-	-
positions[1]	0	1	3	5	6	7	8	-
positions[2]	-	0	1	3	5	6	7	8
positions[3]	-	-	0	1	3	5	6	7

【図 12】



[positions[0],m] = (positions[0], 2, 3, 5, -, -, -, -, -)

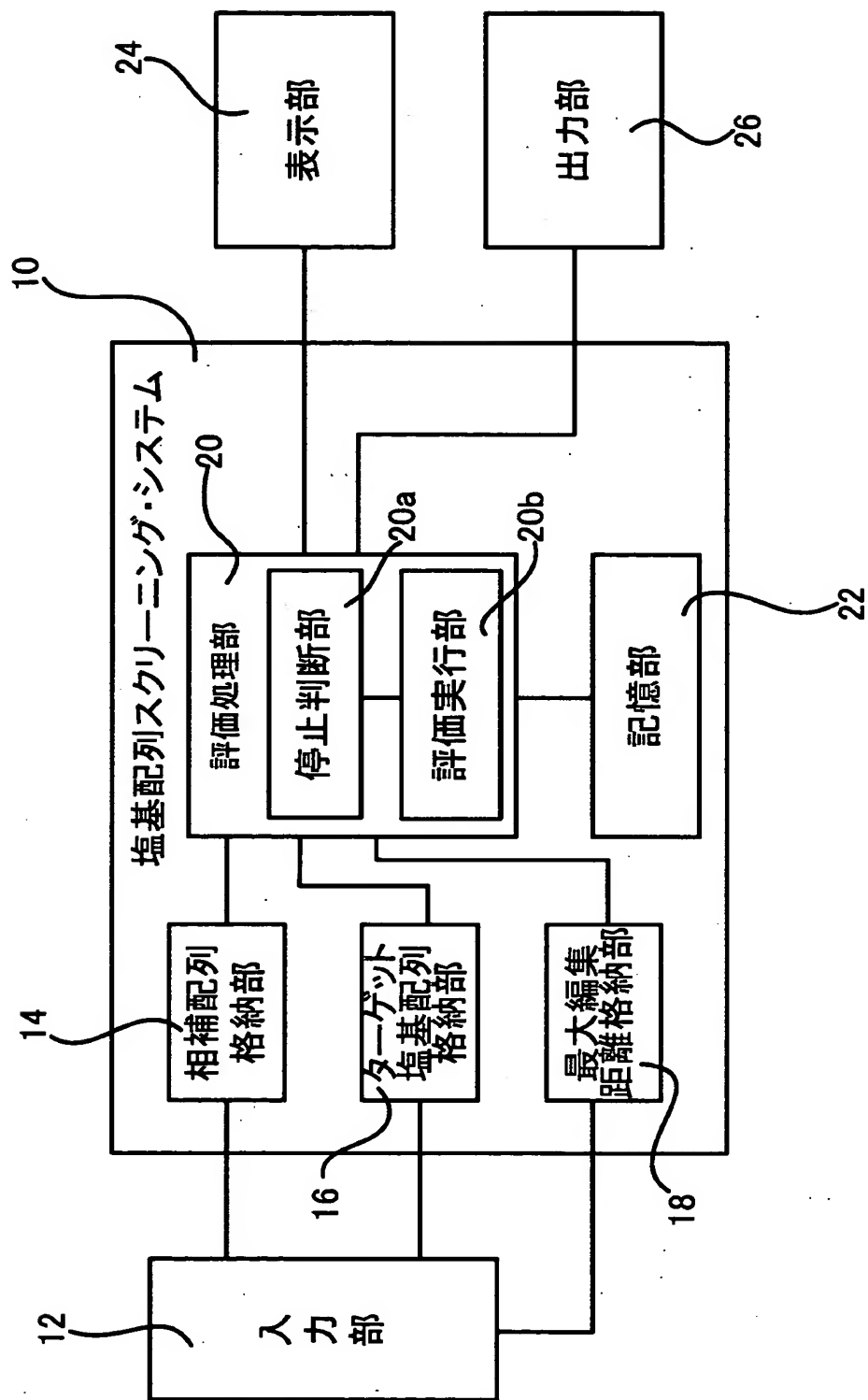
(a)



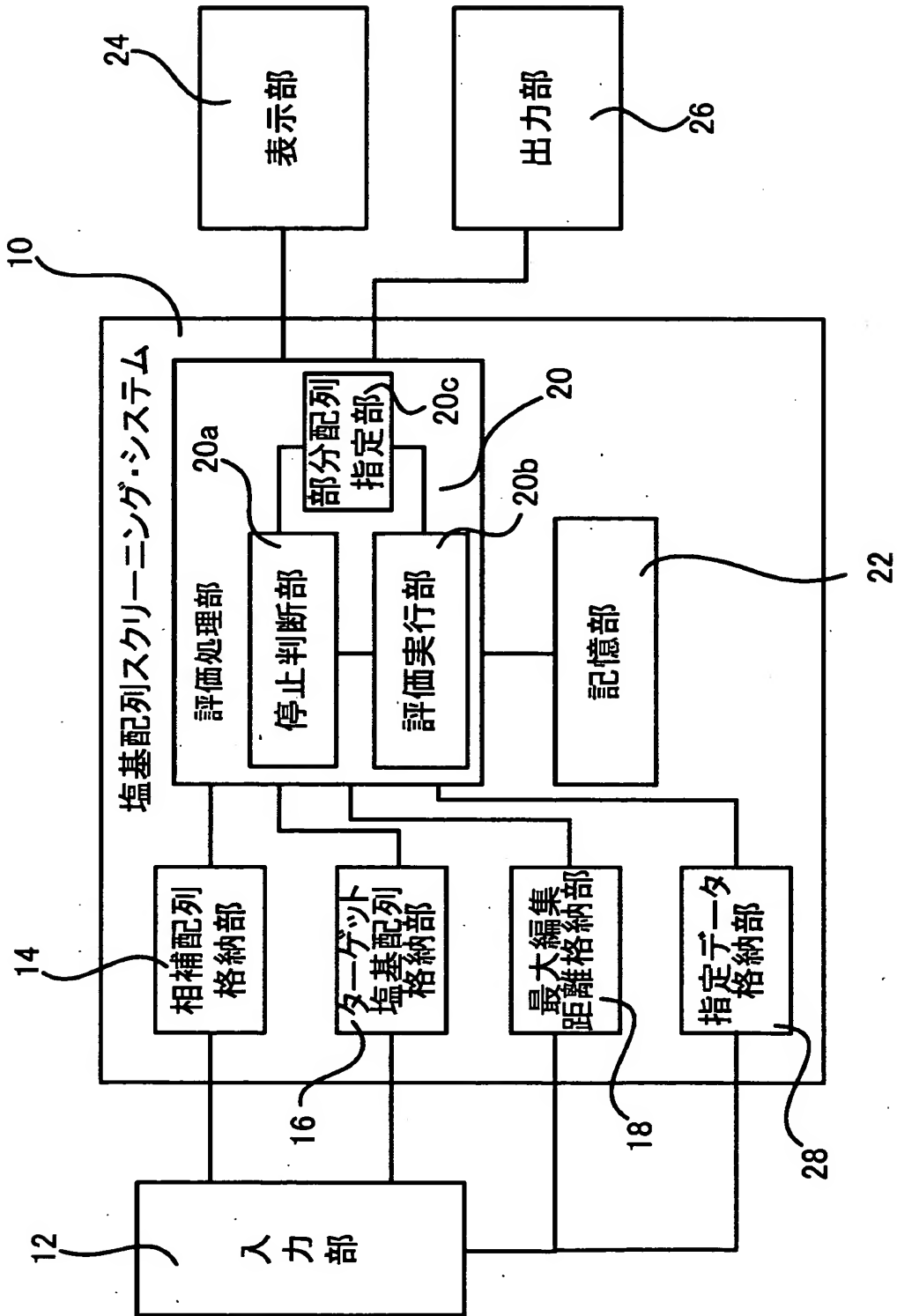
[positions[1],m] = (positions[1], 1, 3, 5, 6, 7, 8)

(b)

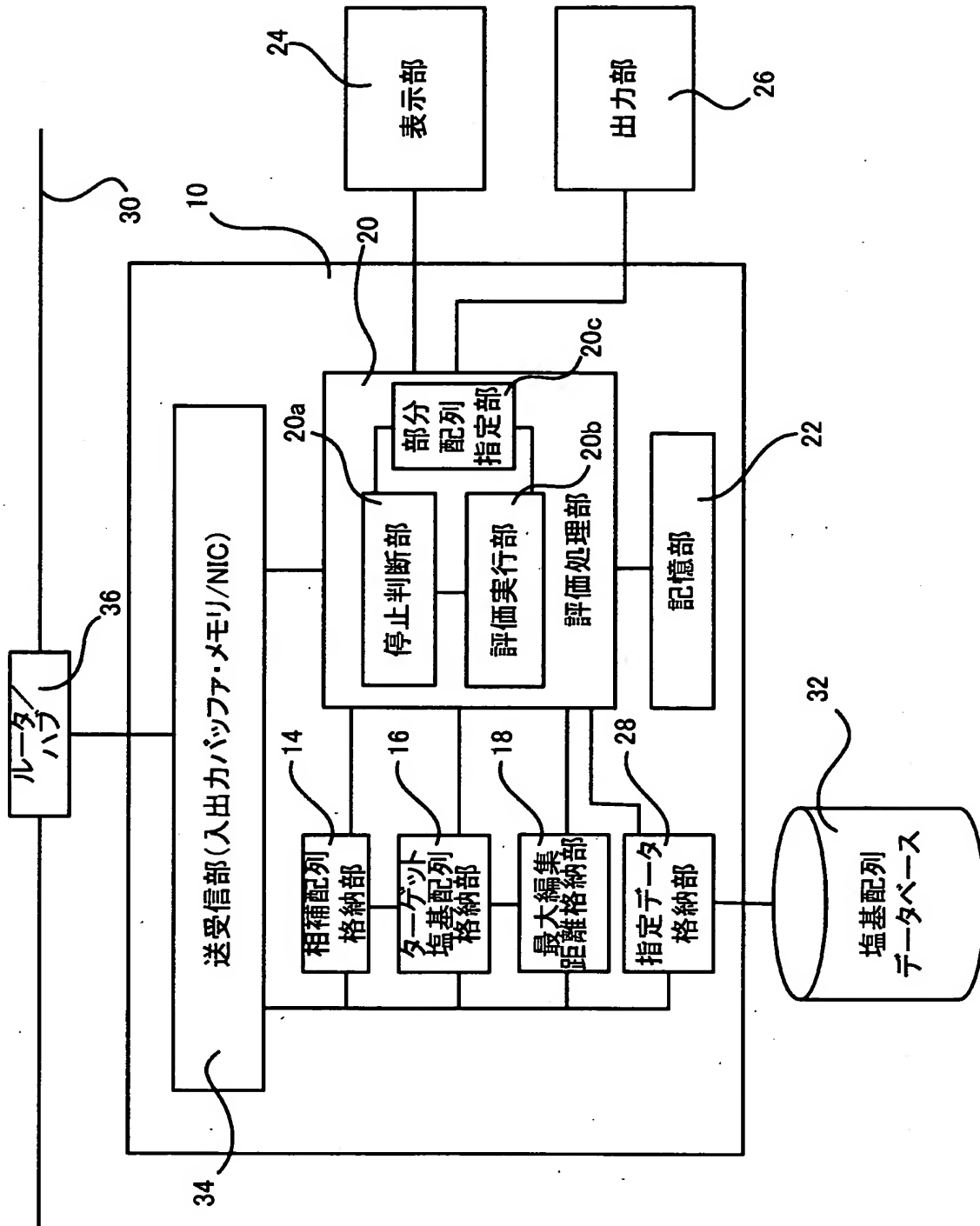
【図 13】



【図 14】



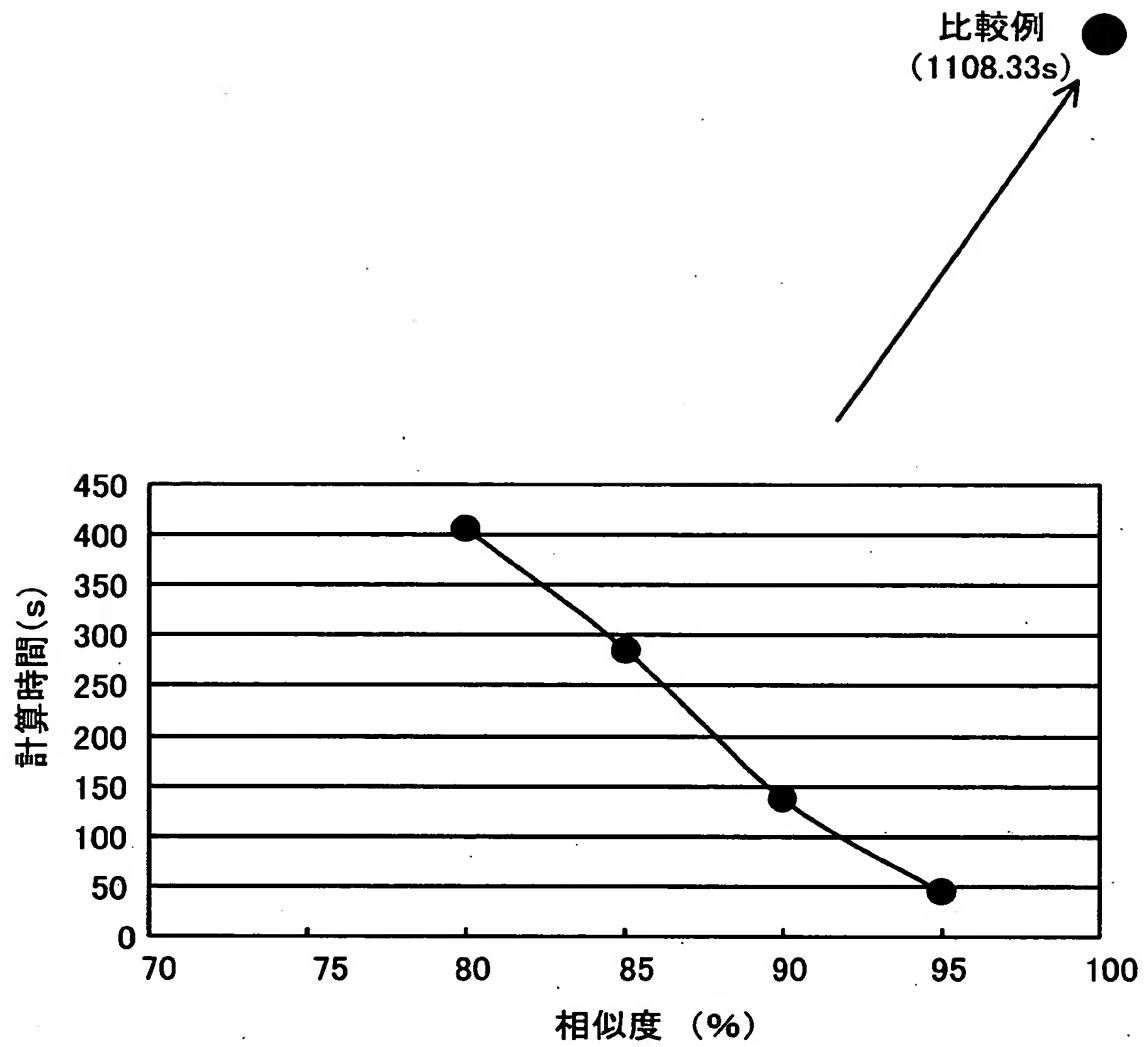
【図 15】



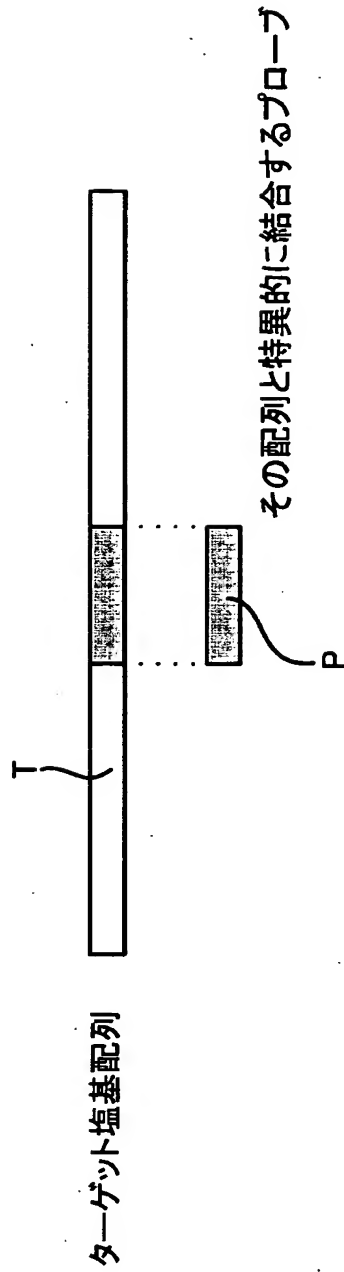
【図 16】

実 施 例				比 較 例 Smith-Waterman
相 似 度				
0.80	0.85	0.90	0.95	-----
405.47s	284.12s	136.20s	45.05s	1108.33s

【図 1 7】



【図 1 8】



【書類名】 要約書

【要約】

【課題】 塩基配列をスクリーニングするためのシステム、そのための方法、プログラム、記録媒体および塩基配列のスクリーニングを実行させるためのサーバを提供する。

【解決手段】 ターゲット塩基配列データおよびプローブ塩基配列とを格納するステップ（S 1 0）と、ターゲット塩基配列に対して結合する可能性のあるプローブ塩基配列から相補配列データを生成し、ターゲット塩基配列とプローブ塩基配列との間に許容する最大の編集距離を格納するステップ（S 1 2）と、各塩基配列データおよび最大編集距離を各格納部から読み出してターゲット塩基配列データと相補配列データとを編集距離の大きな順に結合可能性を評価するステップ（S 1 4）と、結合性の評価の結果を記憶部に格納するステップ（S 1 6）とを含む。

【選択図】 図 3

認定・付加情報

特許出願の番号	特願 2003-050916
受付番号	50300319553
書類名	特許願
担当官	伊藤 雅美 2132
作成日	平成15年 4月 9日

<認定情報・付加情報>

【特許出願人】

【識別番号】	390009531
【住所又は居所】	アメリカ合衆国10504、ニューヨーク州 アーモンク ニュー オーチャード ロード
【氏名又は名称】	インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】	100086243
【住所又は居所】	神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 大和事業所内
【氏名又は名称】	坂口 博

【代理人】

【識別番号】	100091568
【住所又は居所】	神奈川県大和市下鶴間1623番地14 日本アイ・ビー・エム株式会社 大和事業所内
【氏名又は名称】	市位 嘉宏

【代理人】

【識別番号】	100108501
【住所又は居所】	神奈川県大和市下鶴間1623番14 日本アイ・ビー・エム株式会社 知的所有権
【氏名又は名称】	上野 剛史

【復代理人】

【識別番号】	100110607
【住所又は居所】	神奈川県大和市中心林間3丁目4番4号 サクライビル4階 間山国際特許事務所
【氏名又は名称】	間山 進也

次頁無

出 願 人 履 歴 情 報

識別番号 [390009531]

1. 変更年月日 2002年 6月 3日

[変更理由] 住所変更

住 所 アメリカ合衆国10504、ニューヨーク州 アーモンク ニ
ュー オーチャード ロード

氏 名 インターナショナル・ビジネス・マシーンズ・コーポレーショ
ン